On the Convergence of Deep Networks with Sample Quadratic Overparameterization

Asaf Noy, Yi Xu, Yonathan Aflalo, Rong Jin

Machine Intelligence Technology, Alibaba Group

Presenter: Yushun Zhang

February 9, 2021

Introduction

3 Key quesions in deep learning theory:

- Can we reach global min w.h.p? optimization
- Empirical Loss (global min)=0? representation
- Can we generalize well on i.i.d dataset? generalization

Here, we focus on the **global convergence problem:**
given the representation power of DNN:

- When can we reach empirical ϵ-loss via Gradient Descent? (width, depth)
- How many iterations are required?

Introduction

Problem Setup:

- **Assumption 1:** (non-degenerate input). Every two distinct examples x_i, x_j satisfy $\left\|x_i^\top x_j\right\| \le \delta$.
- **Assumption 2:** (common regression labels). Labels are bounded: $\max_i |y_i| \le \frac{m}{d_x}$.
- **Training set:** $\mathcal{T} = \{(x_i, y_i = \Phi_i x_i)\}_{i\in[n}, \ \|x_i\| = 1$
- **NN structure:** L hidden layers, m neurons on each layer
- **Loss:** $\ell\left(W_t\right) = \frac{1}{2}\sum_{i=1}^{n}\left\|f_{W_t}(x_i) - \Phi_i x_i\right\|^2$

Outline

Main result

- **Theorem 1:** Suppose a deep neural network of depth $L = \Omega(\log n)$ is trained by gradient-descent with learning rate $\eta = \frac{d_x}{n^4 L^3 d_y}$, with a width that satisfies,

$$m = \tilde{\Omega}\left(n^2 L d_y\right)$$

Then, with probability of at least $1 - \exp(-\Omega(\sqrt{m}))$ over the random initialization, it reaches ϵ -error within a number of iterations

$$T = O\left(\log\left(\frac{n^3 L}{d_x \epsilon}\right)\right)$$

SOTA results:

Table 1 Comparison of leading works on overparameterized deep nonlinear neural networks trained with gradient-descent.

Work	$\tilde{\Omega}$ (#Neurons)	O_ε (#Iters)	O (Prob)	$\tilde{\Theta}$ (Step)	Remarks
Du[20]	$\frac{n^6}{\lambda_0^4 p^3}$	$\frac{1}{\eta\lambda_0}\log\frac{1}{\varepsilon}$	p	$\frac{\lambda_0}{n^2}$	$\lambda_0^{-1} = \mathrm{poly}\left(e^L, n\right)$, binary -class, smooth activation
Zou[81]	$n^{26}L^{38}$	n^8L^9	–	$\frac{1}{n^{29}L^{47}}$	binary-classification
Allen-Zhu[1]	$n^{24}L^{12}$	$n^6L^2\log(\frac{1}{\varepsilon})$	$e^{-\log^2 m}$	$\frac{1}{n^{28}\log^5 mL^{14}}$	$\propto \mathrm{Poly}(\max_i \lvert y_i \rvert)$
Zou[82]	n^8L^{12}	$n^2L^2\log(\frac{1}{\varepsilon})$	n^{-1}	$\frac{1}{n^8L^{14}}$	–
Ours	$\mathbf{n^2L}$	$\log\left(\frac{\mathbf{n^3L}}{\mathbf{d_x}\epsilon}\right)$	$\mathbf{e^{-\sqrt{m}}}$	$\frac{d_x}{n^4L^3d_y}$	$L = \Omega(\log n)$

Main method

They consider a special structure of Gated-Relu:

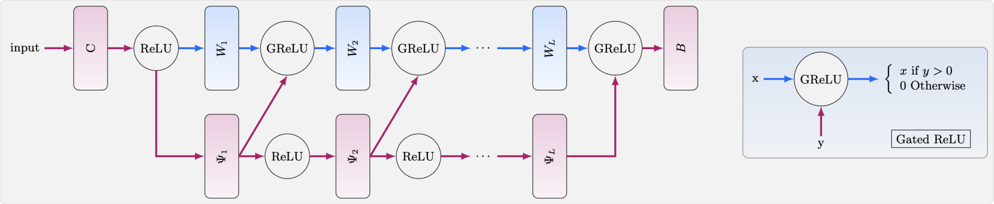

Figure 1: An illustration of the proposed network. Blue layers are trained while red layers set the activations and remain unchanged during training.

initialization: Trained, fixed.

$[W_k]_{i,j} \sim \mathcal{N}(0, 2/m)$, $[\Psi_k]_{i,j} \sim \mathcal{N}(0, 2/m)$, $[C]_{i,j} \sim \mathcal{N}\left(0, 2/d_x\right)$, $[B]_{i,j} \sim \mathcal{N}\left(0, 2/d_y\right)$

Main method

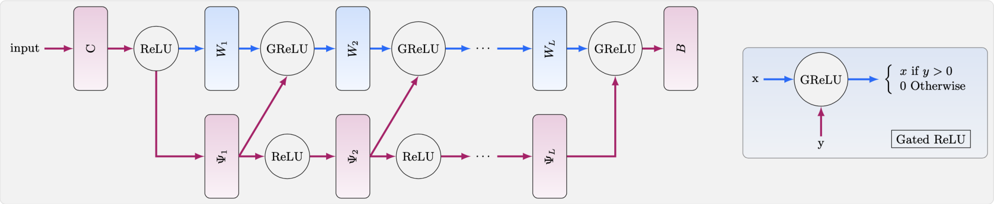

Figure 1: An illustration of the proposed network. Blue layers are trained while red layers set the activations and remain unchanged during training.

- Relu: $f^t(x) = W_2^t D^t W_1^t x$, $D^t = \text{diag}(W_1^t x)_+$, where $(z)_+ = \mathbf{1}_{z>0}$
- As for Relu, D^t is varying along training.

Main method

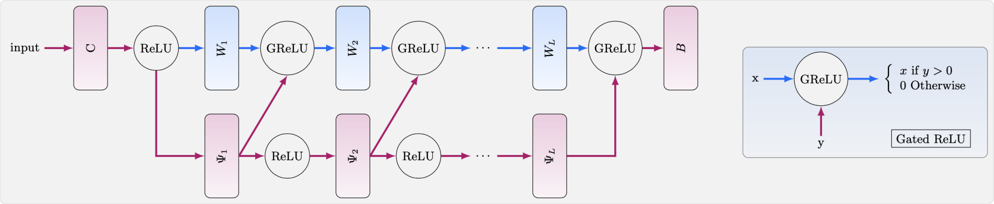

Figure 1: An illustration of the proposed network. Blue layers are trained while red layers set the activations and remain unchanged during training.

- $z_0^i = [Cx_i]^+, \quad z_k^i = \left[\Psi_k z_{k-1}^i\right]^+, \quad D_k^i = \mathrm{diag}\left(\left[z_k^i\right]_+\right) \quad k = 1, \ldots, L$
- GRelu: $f^t(x_i) = W_t^i x_i := BD_L^i W_{t,L} \ldots D_k^i W_{t,k} D_{k-1}^i \ldots D_1^i W_{t,1} D_0^i C x_i$

Main method

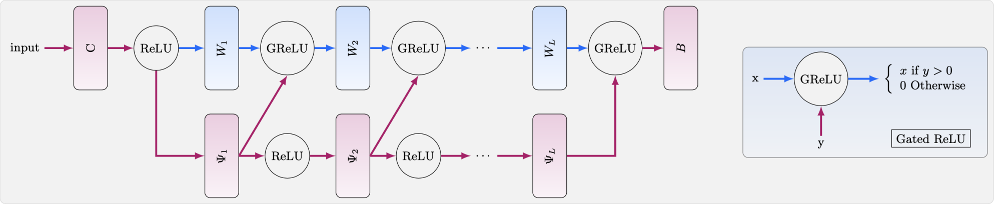

Figure 1: An illustration of the proposed network. Blue layers are trained while red layers set the activations and remain unchanged during training.

- GRelu: $f^t(x_i) = W_t^i x_i := BD_L^i W_{t,L} \dots D_k^i W_{t,k} D_{k-1}^i \dots D_1^i W_{t,1} D_0^i C x_i$
- Here, D_j^i are fixed along training., they are different for different samples.
- For each sample, the active & dead entries are determined from the very beginning, and then fixed.
- The author call it **'fixed activation pattern'**.

Main method

Question 1: what is the benefit of 'fixed activation pattern'?

- Intuitively: adds more 'linearity' to $f(x)$, easier to optimize.
- Technically: makes it easier to bound $W_i^{t+1} - W_i^t$ and $l(W^{t+1}) - l(W^t)$

Question 2: Why is it reasonable to work on a new NN structure Grelu instead of Relu?

- **Theorem 2:** For any Grelu NN, there exists a unique equivalent Relu NN of the same size.
- In practice, people use Resnet in replace of FCN, so it is also legal to modify Relu.

Question 3: How to verify the generalization ability of Grelu?

- Use the equivalence with Relu.

Outline

Proof sketch

We start with the gradient of $\ell\left(W_t\right)$ over $W_{t,k}$, using GRelu, we have:

$$\nabla_k \ell\left(W_t\right) = \sum_{i=1}^{n} \left[F^i_{t,k+1}\right]^\top \left(W^i_t - \Phi_i\right) x_i x_i^\top \left[G^i_{t,k-1}\right]^\top \in \mathrm{R}^{m \times m} \tag{1}$$

where

- $W^i_t := {\color{blue}B D^i_L W_{t,L} \dots D^i_k} W_{t,k} {\color{red}D^i_{k-1} \dots D^i_1 W_{t,1} D^i_0 C} \in \mathrm{R}^{d_y \times d_x}$
- ${\color{blue}F^i_{t,k+1} = B D^i_L W_{t,L} \dots D^i_{k+1} W_{t,k+1} D^i_k} \in \mathrm{R}^{d_y \times m}$
- ${\color{red}G^i_{t,k-1} = D^i_{k-1} W_{t,k-1} \dots D^i_1 W_{t,1} D^i_0 C} \in \mathrm{R}^{m \times d_x}$
- $W^i_t = {\color{blue}F^i_{t,k+1}} W_{t,k} {\color{red}G^i_{t,k-1}}$

Outline

Lemma 1

- **Lemma 1: (Decomposition)** $W_{t+1}^{i} - W_{t}^{i} =$
  $-\eta \sum_{k=1}^{L} F_{t,k+1}^{i} \left[F_{t,k+1}^{i}\right]^{\top} \left(W_{t}^{i} - \Phi_{i}\right) x_{i} x_{i}^{\top} \left[G_{t,k-1}^{i}\right]^{\top} G_{t,k-1}^{i} - \eta \Gamma_{t,i} + \eta^{2} \Delta_{t,i}$
- where $\Gamma_{t,i} := \sum_{k=1}^{L} \sum_{j \neq i} F_{t,k+1}^{i} \left[F_{t,k+1}^{j}\right]^{\top} \left(W_{t}^{j} - \Phi_{j}\right) x_{j} x_{j}^{\top} \left[G_{t,k-1}^{j}\right]^{\top} G_{t,k-1}^{i}$
- $\Delta_{t,i} :=$
  $\sum_{s=2}^{L} (-\eta)^{s-2} \sum_{L \geq k_1 > k_2 \ldots > k_s \geq 1} F_{t,k_1+1}^{i} \nabla_{k_1} \ell\left(W_t\right) D_{k_1-1}^{i} W_{t,k_1-1} \ldots D_{k_s}^{i} \nabla_{k_s} \ell\left(W_t\right) G_{t,k_s-1}^{i}$
- Lemma 1 used 'fixed activation pattern' in Grelu. ($\nabla_k \ell\left(W_t\right)$ is used).
- No assumption on width so far.

Proof of Lemma 1

Proof of lemma 1:

$$W_{t+1}^i - W_t^i = BD_L^i W_{t+1,L} \dots D_1^i W_{t+1,1} D_0^i C - BD_L^i W_{t,L} \dots D_1^i W_{t,1} D_0^i C$$

$$\overset{(a)}{=} BD_L^i \left(W_{t,L} - \eta \nabla_L \ell \left(W_t\right)\right) \dots D_1^i \left(W_{t,1} - \eta \nabla_1 \ell \left(W_t\right)\right) D_0^i C - BD_L^i W_{t,L} \dots D_1^i W_{t,1} D_0^i C$$

$$\overset{(b)}{=} \eta^2 \Delta_{t,i} - \eta \underbrace{\sum_{k=1}^{L} BD_L^i W_{t,L} \dots D_k^i \nabla_k \ell \left(W_t\right) D_{k-1}^i W_{t,k-1} \dots D_1^i W_{t,1} D_0^i C}_{:= Z_t^i}$$

(a): $W_{t+1,k} = W_{t,k} - \eta \nabla_k \ell \left(W_t\right)$ **for any** $k \in [L]$; **(b): uses the definition of** $\Delta_{t,i}$

- Now, simply Z_t^i as:

  $Z_t^i = \sum_{k=1}^{L} F_{t,k+1}^i \nabla_k \ell \left(W_t\right) G_{t,k-1}^i$

  $= \sum_{k=1}^{L} F_{t,k+1}^i \left[F_{t,k+1}^i\right]^\top \left(W_t^i - \Phi_i\right) x_i x_i^\top \left[G_{t,k-1}^i\right]^\top G_{t,k-1}^i + \Gamma_{t,i}$
- We complete the proof by plugging it back.

Outline

Lemma 2

Based on lemma 1 (the change of W_t), we have lemma 2 (the change of loss):

- **Lemma 2:** For any set of positive numbers $a_1, \ldots, a_n$, we have:

$$\ell(W_{t+1}) - \ell(W_t) \leq \sum_{i=1}^{n} \frac{\Lambda_i + \eta^2 a_i}{2} \left\| \left(W_t^i - \Phi_i\right) x_i \right\|^2 + \sum_{i=1}^{n} \frac{\eta^2\left(3\eta^2 + 1/a_i\right)}{2} \left\|\Delta_{t,i} x_i\right\|^2 \quad (2)$$

- where

$$\begin{aligned}
\Lambda_i = & -2\eta \sum_{k=1}^{L} \lambda_{\min}\left(F_{t,k+1}^i \left[F_{t,k+1}^i\right]^\top\right) \lambda_{\min}\left(\left[G_{t,k-1}^i\right]^\top G_{t,k-1}^i\right) \\
& + 2\eta \sum_{k=1}^{L} \sum_{j \neq i} \left|\left\langle G_{t,k-1}^j x_j, G_{t,k-1}^i x_i \right\rangle\right| \left\| F_{t,k+1}^j \left[F_{t,k+1}^i\right]^\top \right\|_2 \\
& + 3\eta^2 L \sum_{k=1}^{L} \lambda_{\max}\left(\mathbf{F_{t,k+1}^i} \left[\mathbf{F_{t,k+1}^i}\right]^\top\right) \left\|\mathbf{G_{t,k-1}^i x_i}\right\|^4 \\
& + 3\eta^2 n L \sum_{k=1}^{L} \sum_{j \neq i} \left|\left\langle G_{t,k-1}^j x_j, G_{t,k-1}^i x_i \right\rangle\right|^2 \left\| F_{t,k+1}^j \left[F_{t,k+1}^i\right]^\top \right\|_2^2
\end{aligned}$$

Lemma 2

- **Lemma 2:** For any set of positive numbers $a_1, \ldots, a_n$, we have:

$$\ell\left(W_{t+1}\right)-\ell\left(W_t\right) \leq \sum_{i=1}^{n} \frac{\Lambda_i+\eta^2 a_i}{2}\left\|\left(W_t^i-\Phi_i\right) x_i\right\|^2+\sum_{i=1}^{n} \frac{\eta^2\left(3\eta^2+1/a_i\right)}{2}\left\|\Delta_{t,i} x_i\right\|^2 \quad (3)$$

- A negative $\Lambda_i+\eta^2 a_i$ values can lead to an linear rate convergence: $\ell\left(W_{t+1}\right)=(1-|\rho|)\ell\left(W_t\right)$
- We wish to bound Λ_i with a negative value as possible.
- Up to now, there is no assumption on width.
- Decompositions are based on 'fixed activation pattern'.

Proof of lemma 2

Proof of lemma 2:

$$\begin{aligned}\ell\left(W_{t+1}\right)-\ell\left(W_t\right) &= \sum_{i=1}^{n}\left\{\frac{1}{2}\left\|\left(W_{t+1}^i-\Phi_i\right)x_i\right\|^2-\frac{1}{2}\left\|\left(W_t^i-\Phi_i\right)x_i\right\|^2\right\}\\ &= \sum_{i=1}^{n}\left\{\left\langle\left(W_{t+1}^i-W_t^i\right)x_i,\left(W_t^i-\Phi_i\right)x_i\right\rangle+\frac{1}{2}\left\|\left(W_t^i-W_{t+1}^i\right)x_i\right\|^2\right\}\end{aligned}$$

Then bound both term respectively using (5 pages of heavy calculation):

- lemma 1.
- $\mathrm{Tr}(AB)=\mathrm{Tr}(BA)$, $\mathrm{Tr}(AB)=\mathrm{vec}^\top(B)\,\mathrm{vec}\left(A^\top\right)$, $\mathrm{vec}(AXB)=B^\top\otimes A\,\mathrm{vec}(X)$, $\langle x,Ay\rangle\le\|x\|\|Ay\|\le\|x\|\|A\|_2\|y\|\le\frac{1}{2}\|A\|_2\left(\|x\|^2+\|y\|^2\right)$
- Young's inequality with $a_i>0$: $\left\langle\Delta_{t,i},\left(W_t^i-\Phi_i\right)x_i\right\rangle\le\frac{1}{2a_i}\left\|\Delta_{t,i}x_i\right\|^2+\frac{a_i}{2}\left\|\left(W_t^i-\Phi_i\right)x_i\right\|^2$

Outline

Assumptions

Now we further bound Λ_i in lemma 2 with a negative value as possible, we need to have the following assumptions w.h.p: (Proof required)

$$\lambda_{\min}\left(F_{t,k}^{i}\left[F_{t,k}^{i}\right]^{\top}\right) \geq \alpha_y, \quad \lambda_{\min}\left(G_{t,k}^{i}\left[G_{t,k}^{i}\right]^{\top}\right) \geq \alpha_x \tag{4}$$

$$\lambda_{\max}\left(\mathbf{F}_{\mathbf{t},\mathbf{k}}^{\mathbf{i}}\left[\mathbf{F}_{\mathbf{t},\mathbf{k}}^{\mathbf{i}}\right]^{\top}\right) \leq \beta_{\mathbf{y}}, \quad \lambda_{\max}\left(\mathbf{G}_{\mathbf{t},\mathbf{k}}^{\mathbf{i}}\left[\mathbf{G}_{\mathbf{t},\mathbf{k}}^{\mathbf{i}}\right]^{\top}\right) \leq \beta_{\mathbf{x}} \tag{5}$$

$$\left|\left\langle G_{t,k-1}^{j} x_j, G_{t,k-1}^{i} x_i\right\rangle\right| \left\|F_{t,k+1}^{j}\left[F_{t,k+1}^{i}\right]^{\top}\right\|_2 \leq \gamma\beta^2 \tag{6}$$

$$\beta^2\gamma n \leq \frac{\alpha^2}{2} \tag{7}$$

- where $\alpha = \sqrt{\alpha_x\alpha_y}$ and $\beta = \sqrt{\beta_y\beta_x}$
- To reach these assumptions, we need: 1. fixed activation pattern. **2.**$\Omega(n^2L)$ **width.**

Lemma 3

Under these assumptions (4)-(7), and Lemma 1 & 2, we have:

- **Lemma 3:** Set $a_i = \beta^4 L^2$, $\ell(W_t) \leq \ell_0$, and

$$\eta = \min\left(\frac{\alpha^2}{12\beta^2\beta_x L}, \frac{1}{3L}, \frac{\alpha^2}{4\beta^4 L}, \frac{1}{\beta^2 L}, \frac{1}{4\sqrt{2}\sqrt{L}e^{\theta/2}\theta^{-1/2}\beta\sqrt{\ell_0}}, \frac{\alpha^2}{1024ne^{2\theta}\theta^{-2}\beta^2\ell_0}\right) \quad (8)$$

  for any $\theta \in (0, 1/5)$, with a probability $1 - L^2\sqrt{m}\exp(-\theta m/[4L] + 6\sqrt{m})$, we have:

$$\ell(W_{t+1}) - \ell(W_t) \leq -\frac{\eta\alpha^2 L}{2}\ell(W_t) \quad (9)$$

- This is known as a linear-rate convergence with rate $\frac{\eta\alpha^2 L}{2}$.
- Theorem 1 follows directly from lemma 3.

Proof of lemma 3

Proof of lemma 3: Using (4)-(7), by setting $a_i = \beta^4 L^2$:

$$\begin{aligned}
\Lambda_i + \eta^2 a_i &\le -2\eta L\alpha^2 + 2\eta L(n-1)\gamma\beta^2 + 3\eta^2 L^2\beta^2\beta_x + 3\eta^2 L^2 n(n-1)\gamma^2\beta^4 + \eta^2 L^2\beta^4 \\
&\le -2\eta L\alpha^2 - 2\eta L\gamma\beta^2 + \eta L\alpha^2 + 3\eta^2 L^2\beta^2\beta_x - 3\eta^2 L^2 n\gamma^2\beta^4 + \frac{3}{4}\eta^2 L^2\alpha^2 + \eta^2 L^2\beta^4 \\
&\le -\eta L\alpha^2 + 3\eta^2 L^2\beta^2\beta_x + \frac{3}{4}\eta^2 L^2\alpha^2 + \eta^2 L^2\beta^4
\end{aligned}$$

By choosing step size η as $\eta \le \min\left(\frac{\alpha^2}{12\beta^2\beta_x L}, \frac{1}{3L}, \frac{\alpha^2}{4\beta^4 L}, \frac{1}{\beta^2 L}\right)$, we have $\Lambda_i \le -\frac{3\eta\alpha^2 L}{4}$, and

$$\ell\left(W_{t+1}\right) - \ell\left(W_t\right) \le -\frac{3\eta\alpha^2 L}{4}\ell\left(W_t\right) + \frac{2\eta^2}{\beta^4 L^2}\sum_{i=1}^{n}\left\|\Delta_{t,i}x_i\right\|^2 \qquad (10)$$

Proof of lemma 3

- Now we need to bound $\left\|\Delta_{t,i} x_i\right\|^2$. Since $\left\|\Delta_{t,i} x_i\right\| \leq \left\|\Delta_{t,i}\right\|_2 \left\|x_i\right\| = \left\|\Delta_{t,i}\right\|_2$, we want to bound $\left\|\Delta_t^i\right\|_2$.
- Recall $\Delta_{t,i} :=$ $\sum_{s=2}^{L} (-\eta)^{s-2} \sum_{L \geq k_1 > k_2 \ldots > k_s \geq 1} F_{t,k_1+1}^i \nabla_{k_1} \ell\left(W_t\right) D_{k_1-1}^i W_{t,k_1-1} \ldots D_{k_s}^i \nabla_{k_s} \ell\left(W_t\right) G_{t,k_s-1}^i$
- To this end, we first bound $\left\|\nabla_k \ell\left(W_t\right)\right\|_2^2$

Proof of lemma 3

$$
\begin{aligned}
&\left\|\nabla_k \ell\left(W_t\right)\right\|_2^2 \\
=&\left\|\nabla_k \ell\left(W_t\right)^{\top} \nabla_k \ell\left(W_t\right)\right\|_2 \\
=&\left\|\left(\sum_{i=1}^{n}\left[F_{t,k+1}^{i}\right]^{\top}\left(W_t^i-\Phi_i\right) x_i x_i^{\top}\left[G_{t,k-1}^{i}\right]^{\top}\right)^{\top}\left(\sum_{i=1}^{n}\left[F_{t,k+1}^{i}\right]^{\top}\left(W_t^i-\Phi_i\right) x_i x_i^{\top}\left[G_{t,k-1}^{i}\right]^{\top}\right)\right\|_2^{\top} \\
=&\left\|\left(\sum_{i=1}^{n} G_{t,k-1}^{i} x_i x_i^{\top}\left(W_t^i-\Phi_i\right)^{\top} F_{t,k+1}^{i}\right)\left(\sum_{i=1}^{n}\left[F_{t,k+1}^{i}\right]^{\top}\left(W_t^i-\Phi_i\right) x_i x_i^{\top}\left[G_{t,k-1}^{i}\right]^{\top}\right)\right\|_2 \\
\leq&\sum_{i=1}^{n}\left\|G_{t,k-1}^{i} x_i x_i^{\top}\left(W_t^i-\Phi_i\right)^{\top} F_{t,k+1}^{i}\left[F_{t,k+1}^{i}\right]^{\top}\left(W_t^i-\Phi_i\right) x_i x_i^{\top}\left[G_{t,k-1}^{i}\right]^{\top}\right\|_2 \\
&+\sum_{i=1}^{n}\sum_{j\neq i}\left\|G_{t,k-1}^{i} x_i x_i^{\top}\left(W_t^i-\Phi_i\right)^{\top} F_{t,k+1}^{i}\left[F_{t,k+1}^{j}\right]^{\top}\left(W_t^j-\Phi_j\right) x_j x_j^{\top}\left[G_{t,k-1}^{j}\right]^{\top}\right\|_2
\end{aligned}
$$

then bound both red and blue terms.

Proof of lemma 3

$$
\begin{aligned}
&\left\| G_{t,k-1}^{i} x_i x_i^{\top} \left(W_t^i - \Phi_i\right)^{\top} F_{t,k+1}^{i} \left[F_{t,k+1}^{i}\right]^{\top} \left(W_t^i - \Phi_i\right) x_i x_i^{\top} \left[G_{t,k-1}^{i}\right]^{\top} \right\|_2 \\
&\overset{(*)}{\leq} \operatorname{Tr}\left(G_{t,k-1}^{i} x_i x_i^{\top} \left(W_t^i - \Phi_i\right)^{\top} F_{t,k+1}^{i} \left[F_{t,k+1}^{i}\right]^{\top} \left(W_t^i - \Phi_i\right) x_i x_i^{\top} \left[G_{t,k-1}^{i}\right]^{\top} \right) \\
&\overset{(a)}{=} \operatorname{Tr}\left(\left[G_{t,k-1}^{i}\right]^{\top} G_{t,k-1}^{i} x_i x_i^{\top} \left(W_t^i - \Phi_i\right)^{\top} F_{k+1}^{i} \left[F_{k+1}^{i}\right]^{\top} \left(W_t^i - \Phi_i\right) x_i x_i^{\top} \right) \\
&\overset{(b)}{=} \operatorname{vec}^{\top}\left(\left(W_t^i - \Phi_i\right) x_i x_i^{\top}\right) \operatorname{vec}\left(\left(\left[G_{t,k-1}^{i}\right]^{\top} G_{t,k-1}^{i} x_i x_i^{\top} \left(W_t^i - \Phi_i\right)^{\top} F_{k+1}^{i} \left[F_{k+1}^{i}\right]^{\top} \right)^{\top} \right) \\
&\overset{(c)}{=} \operatorname{vec}^{\top}\left(\left(W_t^i - \Phi_i\right) x_i x_i^{\top}\right) \operatorname{vec}\left(F_{k+1}^{i} \left[F_{k+1}^{i}\right]^{\top} \left(W_t^i - \Phi_i\right) x_i x_i^{\top} \left[G_{t,k-1}^{i}\right]^{\top} G_{t,k-1}^{i} \right) \\
&\overset{(d)}{=} \operatorname{vec}^{\top}\left(\left(W_t^i - \Phi_i\right) x_i x_i^{\top}\right) \left(\left[G_{t,k-1}^{i}\right]^{\top} G_{t,k-1} \right) \otimes \left(F_{k+1}^{i} \left[F_{k+1}^{i}\right]^{\top} \right) \operatorname{vec}\left(\left(W_t^i - \Phi_i\right) x_i x_i^{\top}\right) \\
&\leq \left\| \left[G_{t,k-1}^{i}\right]^{\top} G_{t,k-1} \right\|_2 \left\| F_{k+1}^{i} \left[F_{k+1}^{i}\right]^{\top} \right\|_2 \left\| \left(W_t^i - \Phi_i\right) x_i \right\|^2 \overset{(5)}{\leq} \beta^2 \left\| \left(W_t^i - \Phi_i\right) x_i \right\|^2
\end{aligned}
$$

(*): $\left\|A^{\top}A\right\|_2 = \|A\|_2^2 \leq \|A\|_F^2 = \operatorname{tr}\left(A^{\top}A\right)$ **(a):** $\operatorname{Tr}(AB) = \operatorname{Tr}(BA)$; **(b):** $\operatorname{Tr}(AB) = \operatorname{vec}^{\top}(B)\operatorname{vec}\left(A^{\top}\right)$; **(c):** $(AB)^{\top} = B^{\top}A^{\top}$; (d): $\operatorname{vec}(AXB) = B^{\top} \otimes A \operatorname{vec}(X)$

Proof of lemma 3

$$
\begin{aligned}
&\sum_{i=1}^{n}\sum_{j\neq i}\left\|G_{t,k-1}^{i}x_i x_i^{\top}\left(W_t^i-\Phi_i\right)^{\top}F_{t,k+1}^{i}\left[F_{t,k+1}^{j}\right]^{\top}\left(W_t^j-\Phi_j\right)x_j x_j^{\top}\left[G_{t,k-1}^{j}\right]^{\top}\right\|_2\\
&\overset{(a)}{=}\sum_{i=1}^{n}\sum_{j\neq i}\left|x_i^{\top}\left(W_t^i-\Phi_i\right)^{\top}F_{t,k+1}^{i}\left[F_{t,k+1}^{j}\right]^{\top}\left(W_t^j-\Phi_j\right)x_j\right|\left\|G_{t,k-1}^{i}x_i x_j^{\top}\left[G_{t,k-1}^{j}\right]^{\top}\right\|_2\\
&\overset{(b)}{\leq}\sum_{i=1}^{n}\sum_{j\neq i}\left\|\left(W_t^i-\Phi_i\right)x_i\right\|\left\|F_{t,k+1}^{i}\left[F_{t,k+1}^{j}\right]^{\top}\right\|_2\left\|\left(W_t^j-\Phi_j\right)x_j\right\|\left\|G_{t,k-1}^{i}x_i\right\|\left\|G_{t,k-1}^{j}x_j\right\|\\
&\overset{(6)}{\leq}\gamma\beta^2\sum_{i=1}^{n}\sum_{j\neq i}\left\|\left(W_t^i-\Phi_i\right)x_i\right\|\left\|\left(W_t^j-\Phi_j\right)x_j\right\|\\
&\overset{(c)}{\leq}\frac{\gamma\beta^2}{2}\sum_{i=1}^{n}\sum_{j\neq i}\left(\left\|\left(W_t^i-\Phi_i\right)x_i\right\|^2+\left\|\left(W_t^j-\Phi_j\right)x_j\right\|^2\right)\\
&\leq n\gamma\beta^2\sum_{i=1}^{n}\left\|\left(W_t^i-\Phi_i\right)x_i\right\|^2=n\gamma\beta^2\ell\left(W_t\right)
\end{aligned}
$$

(a): $\|cA\|_2=|c|\|A\|_2$ **(b):** $\left|x^{\top}Ay\right|\leq\|x\|\|Ay\|\leq\|x\|\|A\|_2\|y\|$ **and** $\left\|xy^{\top}\right\|_2=\left\|x\otimes y^{\top}\right\|_2=\|x\|\|y\|$; **(c) uses Young's inequality.**

Proof of lemma 3

Therefore, combining both red and blue terms, we have:

$$\left\|\nabla_k \ell\left(W_t\right)\right\|_2^2 \leq \left(\beta^2 + n\gamma\beta^2\right) \ell\left(W_t\right) \tag{11}$$

Recall $\Delta_{t,i}$: $\Delta_{t,i} :=$
$\sum_{s=2}^{L}(-\eta)^{s-2} \sum_{L \geq k_1 > k_2 \ldots > k_s \geq 1} F_{t,k_1+1}^{i} \nabla_{k_1} \ell\left(W_t\right) D_{k_1-1}^{i} W_{t,k_1-1} \ldots D_{k_s}^{i} \nabla_{k_s} \ell\left(W_t\right) G_{t,k_s-1}^{i}$
$= \sum_{s=2}^{L}(-\eta)^{s-2} \sum_{L \geq k_1 > k_2 > \ldots > k_s \geq 1} F_{t,k_1+1}^{i} \left(\prod_{\ell=1}^{s} \nabla_{k_\ell} \ell\left(W_t\right) \mathbf{Z}_{\mathbf{k}_\ell-\mathbf{1},\mathbf{k}_{\ell+1}}^{\mathbf{t},\mathbf{i}}\right) G_{t,k_s-1}^{i}$,
where $Z_{k_a,k_b}^{t,i} := D_{k_a}^{i} W_{t,k_a} \ldots W_{t,k_b+1} D_{k_b}^{i}$

- **Lemma 7:**(Not proved yet): for any $\theta \in (0, 1/2)$, with probability $1 - 4L^2 \exp\left(-\theta^2 m / \left[16L^2\right]\right)$, we have: $\left\|\mathbf{Z}_{\mathbf{k}_\mathbf{a},\mathbf{k}_\mathbf{b}}^{\mathbf{t}}\right\|_2 \leq 4\sqrt{L} e^{\theta/2} \theta^{-1/2}$

Proof of lemma 3

w.h.p:

$$\left\|\Delta_{t,i}\right\|_2 \overset{(a)}{\leq}$$
$$\sum_{s=2}^{(a)} \eta^{s-2} \sum_{L\geq k_1>k_2>\ldots>k_s\geq 1} \left\|F^i_{t,k_1+1}\right\|_2 \left\|G^i_{t,k_s-1}\right\|_2 \left(\prod_{\ell=1}^{s} \left\|\nabla_{k_\ell}\ell\left(W_t\right)\right\|_2 \left\|Z^{t,i}_{k_\ell-1,k_{\ell+1}}\right\|_2\right)$$
$$\overset{(b)}{\leq} \sum_{s=2}^{L} \eta^{s-2} \begin{pmatrix} L \\ s \end{pmatrix} \beta \left(\sqrt{\left(\beta^2+n\gamma\beta^2\right)\ell\left(W_t\right)} \times 4\sqrt{L}e^{\theta/2}\theta^{-1/2}\right)^s$$
$$\overset{(c)}{\leq} \sum_{s=2}^{L} \beta\eta^{s-2} \left(2\sqrt{2}\sqrt{L}e^{\theta/2}\theta^{-1/2}\beta\sqrt{\ell\left(W_t\right)}\right)^s$$
$$\overset{(d)}{\leq} 8Le^{\theta}\theta^{-1}\beta^3\ell\left(W_t\right)\sum_{s=0}^{L-2}\eta^s \left(2\sqrt{2}\sqrt{L}e^{\theta/2}\theta^{-1/2}\beta\sqrt{\ell_0}\right)^s \quad \overset{(e)}{\leq} \frac{8Le^{\theta}\theta^{-1}\beta^3\ell(W_t)}{1-2\sqrt{2}\eta\sqrt{L}e^{\theta/2}\theta^{-1/2}\beta\sqrt{\ell_0}}$$

(a): $\|AB\|_2 \leq \|A\|_2\|B\|_2$ **and** $\|A+B\|_2 \leq \|A\|_2 + \|B\|_2$, **(b): lemma 7,**
(c): $\binom{L}{s} = \frac{L!}{(L-s)!s!} \leq \frac{L!}{(L-s)!} = L(L-1)\ldots(L-s+1) \leq L^s$,**(d):** $\ell\left(W_t\right) \leq \ell_0$, **(e): choose**
$\eta < 1/\left(2\sqrt{2}\sqrt{L}e^{\theta/2}\theta^{-1/2}\beta\sqrt{\ell_0}\right)$

Proof of lemma 3

w.h.p:

$$\ell\left(W_{t+1}\right)-\ell\left(W_t\right) \leq -\frac{3\eta\alpha^2 L}{4}\ell\left(W_t\right)+\frac{128n\eta^2 e^{2\theta}\theta^{-2}\beta^2\ell^2\left(W_t\right)}{1-2\sqrt{2}\eta\sqrt{L}e^{\theta/2}\theta^{-1/2}\beta\sqrt{\ell_0}} \tag{12}$$

Choose $\eta \leq \frac{\alpha^2}{1024ne^{2\theta}\theta^{-2}\beta^2\ell_0}$, or in summary:
$\eta = \min\left(\frac{\alpha^2}{12\beta^2\beta_x L}, \frac{1}{3L}, \frac{\alpha^2}{4\beta^4 L}, \frac{1}{\beta^2 L}, \frac{1}{4\sqrt{2}\sqrt{L}e^{\theta/2}\theta^{-1/2}\beta\sqrt{\ell_0}}, \frac{\alpha^2}{1024ne^{2\theta}\theta^{-2}\beta^2\ell_0}\right)$:
we have

$$\ell\left(W_{t+1}\right)-\ell\left(W_t\right) \leq -\frac{\eta\alpha^2 L}{2}\ell\left(W_t\right) \tag{13}$$

- Lemma 3 proof completed: linear convergence.
- but Lemma 3 is based on assumptions (4)-(7): how to prove them? Requires width of $\Omega(n^2)$.

Outline

Assumptions

with high probability, with $\alpha = \sqrt{\alpha_x \alpha_y}$ and $\beta = \sqrt{\beta_y \beta_x}$:

$$\lambda_{\min}\left(F^i_{t,k}\left[F^i_{t,k}\right]^\top\right) \geq \alpha_y, \quad \lambda_{\min}\left(G^i_{t,k}\left[G^i_{t,k}\right]^\top\right) \geq \alpha_x \tag{14}$$

$$\lambda_{\max}\left(\mathbf{F}^{\mathbf{i}}_{\mathbf{t,k}}\left[\mathbf{F}^{\mathbf{i}}_{\mathbf{t,k}}\right]^\top\right) \leq \beta_{\mathbf{y}}, \quad \lambda_{\max}\left(\mathbf{G}^{\mathbf{i}}_{\mathbf{t,k}}\left[\mathbf{G}^{\mathbf{i}}_{\mathbf{t,k}}\right]^\top\right) \leq \beta_{\mathbf{x}} \tag{15}$$

$$\left|\left\langle G^j_{t,k-1}x_j, G^i_{t,k-1}x_i\right\rangle\right| \left\|F^j_{t,k+1}\left[F^i_{t,k+1}\right]^\top\right\|_2 \leq \gamma\beta^2 \tag{16}$$

$$\beta^2\gamma n \leq \frac{\alpha^2}{2} \tag{17}$$

- Under these assumptions, we have shown the linear convergence in lemma 3.
- When do these assumptions hold ? we need: 1. fixed activation pattern. **2.**$O(n^2)$ **width.**

Lemma 5

To prove them, we need **lemma 5.**

- **Lemma 5:** With a probability at least $1 - 2L \exp\left(-\theta^2 m / \left[16L^2\right]\right)$, for any $\theta \in (0, 1/2)$, we have:

$$\left\| Z^{1,i}_{k_a,k_b} \right\|_2 \leq \sqrt{12L} e^{\theta/2} \theta^{-1/2} \tag{18}$$

and with a probability $1 - 4L^2 \exp\left(-\theta^2 m / \left[8L^2\right] + 3d_x\right)$, we have:

$$\left\| Z^{1,i}_{k_a,k_b} C \right\|_2 \leq \sqrt{\frac{3m}{d_x}} e^{\theta/2} \tag{19}$$

- where $Z^{t,i}_{k_a,k_b} := D^i_{k_a} W_{t,k_a} \dots W_{t,k_b+1} D^i_{k_b}$.
- **Proof:** covering number, concentration bound for chi-square distribution. **fixed activation pattern.** (treating D as a constant matrix.)

Proof of Assumptions (4) (5)

We now prove (4) and (5): with high probability:

$$\lambda_{\min}\left(F^i_{t,k}\left[F^i_{t,k}\right]^\top\right) \geq \alpha_y, \quad \lambda_{\min}\left(G^i_{t,k}\left[G^i_{t,k}\right]^\top\right) \geq \alpha_x$$

$$\lambda_{\max}\left(\mathbf{F}^{\mathbf{i}}_{\mathbf{t,k}}\left[\mathbf{F}^{\mathbf{i}}_{\mathbf{t,k}}\right]^\top\right) \leq \beta_{\mathbf{y}} = \frac{\mathbf{27m}}{\mathbf{4d_y}}, \quad \lambda_{\max}\left(\mathbf{G}^{\mathbf{i}}_{\mathbf{t,k}}\left[\mathbf{G}^{\mathbf{i}}_{\mathbf{t,k}}\right]^\top\right) \leq \beta_{\mathbf{x}} = \frac{\mathbf{27m}}{\mathbf{4d_x}}$$

Proof: Define $\delta W_{t,k} := W_{t,k} - W_{1,k}$, we prove (5) ((4) is Similarly):

$$\begin{aligned}
G^i_{t,k} =& D^i_k\left(W_{1,k} + \delta W_{t,k}\right)\dots D^i_1\left(W_{1,1} + \delta W_{t,1}\right) D^i_0 C \\
=& D^i_k W_{1,k}\dots D^i_0 C \\
&+ \sum_{s=1}^{k}\sum_{k_1>k_2>\dots>k_s} D^i_k W_{1,k}\dots D^i_{k_1}\delta W_{t,k_1} D^i_{k_1-1} W_{1,k_1-1}\dots D^i_{k_s}\delta W_{t,k_s} D^i_{k_s-1} W_{1,k_s-1}\dots D^i_0 C \\
=& Z^{1,i}_{k,0} C + \sum_{s=1}^{k}\sum_{k_1>k_2>\dots>k_s}\left(\prod_{j=1}^{s} Z_{k_{j-1},k_j}\delta W_{t,k_j}\right) Z_{k_s-1,0} C.
\end{aligned}$$

Proof of Assumptions (4) (5)

Use Lemma 5:

- $\max_{u\in\mathrm{R}d_x} \frac{\left\|G_{t,k}^i u\right\|}{\|u\|} \leq \sqrt{\frac{3m}{d_x}}e^{\theta/2} + \sqrt{\frac{3m}{d_x}}e^{\theta/2}\sum_{s=1}^{L}\left(L\tau\sqrt{12L}e^{\theta/2}\theta^{-1/2}\right)^s \leq \sqrt{\frac{3m}{d_x}}e^{\theta/2}\left(1+\frac{L\tau\sqrt{12L}e^{\theta/2}\theta^{-1/2}}{1-L\tau\sqrt{12L}e^{\theta/2}\theta^{-1/2}}\right)$
- Choose $\theta \in (0, 1/2)$ which satisfies $L\tau 3\sqrt{12L}e^{2\theta}\theta^{-1/2} \leq \frac{1}{9}$:
  we have, with a probability $1 - 4L^3 \exp\left(-\theta^2 m/\left[16L^2\right] + 3d_x\right)$:

$$\max_{u\in\mathrm{R}^{d_x}} \frac{\left\|G_{t,k}^i u\right\|}{\|u\|} \leq \frac{3}{2}\sqrt{\frac{3m}{d_x}} \tag{20}$$

- Similar analysis applies to $\left\|F_{t,k}^i\right\|_2$, and $\lambda_{min}(\cdot)$
- Up to now, no requirement on width. but we used **fixed activation pattern** in lemma 5.

Proof of Assumptions (6)

As for assumption (6), we have the following theorem:

- **Theorem 6:** With a probability $1-\left(4L^2+n^2\right)\exp\left(-\Omega\left(\sqrt{m}+\max\{d_x,d_y\}\right)\right)$, we have:

$$\left\|F^j_{t,k}\left[F^i_{t,k}\right]^\top\right\|_2 \le C'\left(\frac{1}{m^{1/4}}+\left(\frac{5}{6}\right)^{L-k}+L^{3/2}\tau\right)\beta_y$$

$$\left\langle G^j_{t,k}x_j, G^i_{t,k}x_i\right\rangle | \le C'\left(\frac{1}{m^{1/4}}+\delta\left(\frac{5}{6}\right)^{k}+L^{3/2}\tau\right)\beta_x$$

- Therefore, $\left|\left\langle G^j_{t,k-1}x_j, G^i_{t,k-1}x_i\right\rangle\right|\left\|F^j_{t,k+1}\left[F^i_{t,k+1}\right]^\top\right\|_2 \le \gamma\beta^2$, with

$$\gamma = C''\left(L^3\tau^2+\delta\left(\frac{5}{6}\right)^{L}+\frac{1}{m^{1/2}}\right)$$

- where $\tau := \max_{1\le t\le T}\max_{k\in[L]}\|W_{t,k}-W_{1,k}\|_2$
- **Proof:** Similarly as before, use lemma 5 repeatedly.

Proof of Assumptions (7)

As for Assumption (7), with $\alpha = \sqrt{\alpha_x \alpha_y}$:

$$\beta^2 \gamma n \leq \frac{\alpha^2}{2} \tag{21}$$

Since $\gamma = C'' \left(L^3 \tau^2 + \delta \left(\frac{5}{6}\right)^L + \frac{1}{m^{1/2}} \right)$ in Theorem 6, we must have:

$$\gamma = C'' \left(\left(L^{3/2} \tau\right)^2 + \delta \left(\frac{5}{6}\right)^L + \frac{1}{m^{1/2}} \right) = O\left(\frac{1}{n}\right)$$

- To meet the above condition, we must have:
  $L^{3/2} \tau = O\left(\frac{1}{\sqrt{n}}\right), \quad L = \Omega(\log n), \quad m = \Omega\left(n^2\right)$
- we thus need to bound τ by choosing an appropriate width.

Outline

Theorem 1

After all the suffering, now lets come back to theorem 1:

- **Theorem 1:** Suppose a deep neural network of depth $L = \Omega(\log n)$ is trained by gradient-descent with learning rate $\eta = \frac{d_x}{n^4 L^3 d_y}$, with a width that satisfies,

$$m = \tilde{\Omega}\left(n^2 L d_y\right)$$

  Then, with probability of at least $1 - \exp(-\Omega(\sqrt{m}))$ over the random initialization, it reaches ϵ -error within a number of iterations

$$T = O\left(\log\left(\frac{n^3 L}{d_x \epsilon}\right)\right)$$

Proof of Theorem 1

To meet Assumption (7): $\beta^2\gamma n \leq \frac{\alpha^2}{2}$,

we must have:

- $L^{3/2}\tau = O\left(\frac{1}{\sqrt{n}}\right), \quad L = \Omega(\log n), \quad m = \Omega\left(n^2\right)$
- we thus need to bound $\tau = \max_{1\leq t\leq T} \max_{k\in[L]} \|W_{t,k} - W_{1,k}\|_2$
- Based on linear convergence (lemma 1,2, 3), assume we can get ϵ loss, the number of iterations needed is: $T = \frac{2}{\eta\alpha^2 L}\log\frac{\ell_0}{\epsilon}$.
- On the other hand:
  $\tau \leq \eta\sum_{t=1}^{T-1}\max_{k\in[L]}\|\nabla_k\ell\left(W_t\right)\| \overset{(a)}{\leq} \eta\beta\sum_{t=1}^{T-1}\sqrt{2\ell\left(W_t\right)} \overset{(b)}{\leq} \eta\beta T\sqrt{2\ell_0} = \frac{2\beta\sqrt{2\ell_0}}{\alpha^2 L}\log\frac{\ell_0}{\epsilon}$
- where **(a):**$\|\nabla_k\ell\left(W_t\right)\|_2^2 \leq \left(\beta^2 + n\gamma\beta^2\right)\ell\left(W_t\right)$**, (b):** $\ell\left(W_t\right) \leq \ell_0$
- we need to further bound ℓ_0.

Proof of Theorem 1

Similarly with lemma 5, w.h.p:

$$\ell_0 = \frac{1}{2}\sum^{n} \left\| \left(W_0^i - \Phi_i\right) x_i \right\|^2 \leq \frac{1}{2}\sum^{n} \left(\| \left(W_0^i x_i \right\|^2 + \| y_i \|^2 \right) = \frac{n}{2}\left(\frac{3m}{d_x}e^{\theta} + \frac{m}{d_x}\right) \leq \frac{4mn}{d_x} \tag{22}$$

Where we also used $\theta < 0.5$ and Assumption (2): $\max_i |y_i| \leq \frac{m}{d_x}$, now we have:

$$L^{3/2}\tau = \tilde{O}\left(\frac{\sqrt{\ell_0 d_x d_y L}}{m}\right) = O\left(\frac{1}{\sqrt{n}}\right) \tag{23}$$

Therefore, to meet $L^{3/2}\tau = O\left(\frac{1}{\sqrt{n}}\right)$, we need the width:

$$m = \tilde{\Omega}\left(n^2 L d_y\right)$$

Outline

Proof sketch

The main story line goes like this:

- Regular relu is difficult: $f^t(x) = W_2^t D^t W_1^t x$.
- Grelu: fixed activation pattern, more 'linearity' $\Rightarrow$ Lemma 1, Lemma 2.
- At initialization: Grelu $\Rightarrow$ Lemma 5 repeatedly to bound weight submatrix Z (made up with D)
- lemma 5 $\Rightarrow$ (4), (5), (6) w.h.p .
- With certain width $\Rightarrow$ bounded movement δW, $\Rightarrow$we can get assumption (7)
- (4)(5)(6)(7)$\Rightarrow$ linear convergence Lemma 3.
- Lemma 1,2,3 $\Rightarrow$ global convergence Theorem 1.

Proof sketch

Question: How to verify the generalization ability of Grelu?

- Use the equivalence with Relu.
- **Theorem 2:** Let $W_t = \left(W_{t,1}, \ldots, W_{t,L}; C, B, \Psi_{[L]}\right)$ be an overparameterized Grelu network of depth L and width m, trained by gradient-descent for t steps. Then, a unique equivalent ReLU network of the same sizes $W_t = \left(W'_{t,1}, \ldots, W'_{t,L}; C, B\right)$ can be obtained, with identical intermediate and output values over the train set.
- **Proof idea:** match the output of the Gated ReLU and the input of the ReLU one, i.e., we seek for W'_k, such that, for any sample i:
$W'_k z_{k-1}^{\mathrm{ReLU}^i} = \mathrm{GReLU}\left(W_{t,k} z_{k-1}^i, \Psi_i z_{k-1}^i\right)$
- $W'_k = \left(\mathrm{GReLU}\left(W_{t,k} {z'}_{k-1}, \Psi_i z_{k-1}\right)\right)^{\dagger} z_{k-1}^{\mathrm{ReLU}}$